Measuring short-form factuality in large language models

Year: 2024 | *Citations: 191* | *Authors: Jason Wei, Nguyen Karina, Hyung Won Chung*

Abstract

We present SimpleQA, a benchmark that evaluates the ability of language models to answer short, fact-seeking questions. We prioritized two properties in designing this eval. First, SimpleQA is challenging, as it is adversarially collected against GPT-4 responses. Second, responses are easy to grade, because questions are created such that there exists only a single, indisputable answer. Each answer in SimpleQA is graded as either correct, incorrect, or not attempted. A model with ideal behavior would get as many questions correct as possible while not attempting the questions for which it is not confident it knows the correct answer. SimpleQA is a simple, targeted evaluation for whether models"know what they know,"and our hope is that this benchmark will remain relevant for the next few generations of frontier models. SimpleQA can be found at https://github.com/openai/simple-evals.